J Labor Res (2012) 33:173–195
DOI 10.1007/s12122-011-9124-7

Trends in the Gender Pay Gap in Spain: A Semiparametric Analysis

Ignacio Moral-Arce · Stefan Sperlich · Ana I. Fernández-Saínz · Maria J. Roca

Published online: 22 October 2011

Abstract This article studies the trend in wage discrimination in Spain from 1995 to 2002, when the third plan for equal opportunities for men and women was in action. To account for the criticism of Heckman et al. (J Hum Cap 2:1–31, 2008), we first introduce a novel approach to the analysis of wage discrimination with methods that are robust to model (mis-) specification. Following their idea, we apply semi-parametric methods for the Oaxaca-Blinder decomposition of wage differentials between men and women. We extend the methodology to semiparametric quantile estimation. The study is completed by some descriptive analysis, also based on nonparametric techniques. We find that, while the wage gap has diminished from 1995 to 2002 this is mainly due the smaller gap in returns of endowments for wages above the median, and due to the endowments of women for lower and particularly high wages. Respective the quantiles, in contrast to other EU member states, the Spanish wage gap is widest for low wages but almost U-shaped in 2002 whereas this was not that evident in 1995.

Keywords Wage differentials · Semiparametric regression · Counterfactual quartile regression · Gender pay gap · Oaxaca-Blinder decomposition

JEL Classification C14 · J16

I. Moral-Arce
Institute for Fiscal Studies, Ministry of Finance, Madrid, Spain

S. Sperlich (✉)
Département des Sciences Économiques, Université de Genève, Bd du Pont d'Arve 40, CH-1211 Genève 4, Switzerland
e-mail: stefan.sperlich@unige.ch

A. I. Fernández-Saínz
Departamento de Economía Aplicada III, Universidad del País Vasco, Basque, Spain

M. J. Roca
Cantabrian Chamber of Commerce, Santander, Spain

Introduction

For the past thirty years much attention has been paid to wage discrimination; in the US the focus has been more on discrimination due to race, while in Europe it has been mainly on gender discrimination. This article concentrates on the latter, investigating trends in the gender pay gap in Spain from the nineties to the beginning of the new millennium, when different political activities where undertaken to enforce gender equality in the EU and inside Spain.

The labor market in Spain has changed enormously from economic, legislative and social viewpoints in the last 35 years, since the advent of democracy in 1975. Apart from the transition to democracy, another driving force behind the changes in the labor market in Spain was the country's entry into the European Union in 1986. Changes had to be made in the 80s prior to enter into the EU, and further important changes took place throughout the 90s. These changes affected both societal and economic aspects.

Furthermore, Spain has experienced considerable economic growth over the last thirty years. The economy boomed from 1986 to 1990, averaging 5% annual growth. After a Europe-wide recession in the early 1990s, the Spanish economy resumed moderate growth in 1994, which continued (up to the international financial crisis) thanks to the advantages provided by enormous quantities of economic aid from the EU. Spain has had one of the highest growth rates of any EU country in several years of the studied period. Unemployment in the Spanish labor market in the late 70s was at a level close to full employment. In 1976 the rate was 4.64% for men, and 4.94% for women. Unemployment increased sharply in the 80s, especially among women, and peaked in the early 90s at 20.51% for men (first quarter of 1994) and 31.96% for women (fourth quarter of 1994). The situation began to improve in the second half of the 90s, but to a lesser degree for women. At the end of this period women were in a more precarious situation than men.

Many changes in legislation governing the treatment of men and women were introduced between the 70s and the 90s. There has been existing the general belief that one of the fundamental characteristics of the Spanish labor market was, and maybe is, the persistent and strong wage discrimination due to gender; for a similar job, men are clearly paid more than women. The equality principle was enshrined in Article 14 of the Spanish Constitution of 1978, which clearly prohibited discrimination on grounds of gender. The Workers' Statute Act of 1980 (amended several times since that) contains a number of rules on the equal treatment of men and women. Wage equality for work of equal value was established in Article 28. Moreover, the 3rd Plan for Equal Opportunities for Men and Women (1997–2000) recognized the need to incorporate more women into remunerated labor, the persistence of unjustifiable wage inequalities for women already working, and the existence of large-scale segregation of female employment. To palliate this unequal situation, a number of actions were taken under the Plan to provide women with real access to employment with full social and economic rights by encouraging structural changes and transformations that favored this purpose, with special emphasis on the reconciliation of family and working life. In total, there were four Plans for Equal Opportunities until today, the Activity plans for the Employment (with a special emphasis on gender equality) in 1998, and not to forget the EU Strategy Plan for gender equality from June 7 of 2000.

As well as national policy, there is a large body of European legislation on equal treatment and labor market access of women and men. Among the legislative advances that deserve to be pointed out is the inclusion of the principle of non-discrimination on grounds of gender as one of the objectives of the EU. In 1997 the Member States jointly decided to implement a new strategy for employment in which equal opportunities should be an important and explicit component since it is one of the pillars of the guidelines for employment in the Union. Partly, this has been done quite successfully such that the wage-gap has been converging, i.e. decreasing (especially) in the Mediterranean EU countries which today have a smaller gap than the other members, see OCDE (2002). However, all this has had a positive impact mainly on the discrimination in the public sector where today, one can hardly find a pay-gap anymore in Spain, see Ullibarri (2003) and Aláez et al. (2009).

Another point worth highlighting is that considerable social changes took place over the period under study, such as the emancipation of women and their massive incorporation into the labor market (there was a great increase of the rate of female participation in the labor force among the 25–54 age group, with an employment rate up from less than 30% in 1976 to over 40% in 1995) and also an increase in the level of education of women. Theory says that the societal discrimination in Spain entails statistical discrimination in endowments and experience (or loyalty towards the enterprise) which in turn is used to pay women lower wages especially at the entry and during the first years of professional activity. De la Rica et al. (2008) argue that this would only hold for low educated women but vanish over time, whereas high educated women start with similar entry salaries as men but then reach the so-called glass ceilings which are not faced by low educated women; compare also with del Rio et al. (2011). Over the last 35 years Spanish women have dedicated much more time to studying (at present the percentage of women studying at university is higher than that of men), which is one reason why women enter the labor market later. To mention is the Spanish Law for the Conciliation of Family and Labor from 1999. Another demographic outcome worth describing is that (comparing 1995 to 2002) women marry later, which considerably delays the age at which they have their first child. This has allowed women to achieve some success as a labor force on the one hand, and reduced the fertility rate to the lowest level in the history of Spain, and actually alsoone of the lowest in the world. When the total fertility rate was only 1.36 in 1990, it even went down to 1.17 in 1995 and recovered 1.26 in 2002 but mainly thanks to a strong immigration which contributed with a fertility rate of 2.05. Note that in their review for Europe, Arulampalam et al. (2007) emphasized the need of child care provision to fight the pay gap.

As mentioned above, most of these elements entailed a massive incorporation of women into the labor market. Albeit for different reasons, the market has not been able to absorb this increase of labor force, resulting in an excess of female labor supply although labor participation in the mid nineties in Spain was clearly below the European level. This translates to some degree into a situation of discrimination against women on the labor market: it is known that their unemployment rates are higher than those of men in most sectors, the jobs that women take do not involve the same degree of responsibility or decision-making power as those of men, and women's participation is limited to a few sectors of the economy. This can seriously affect the pay gap as for the Spanish industrial sector of which we know that the wage dispersion has increased dramatically, see for example Fernandez et al. (2000) or Aláez and Ullibarri (2001). Some further contribution in this direction are Garcia et al. (2001), Simón (2010), Simón et al. (2008), Aláez et al. (2003), see also Miller (2009) for an US study. Various papers show that the Spanish wage gap remains notable, mainly in the private sector on which we therefore will concentrate. A different separation has been considerd in De la Rica et al. (2008) which partitioned their population by education. They found (along the wage scale) a decreasing gap for lower educated and increasing one for highly educated women. For a pooled sample this could result in a U-shape on the wage scale, similar to what they showed in their Fig. 1a.

Our approach differs from these articles in several aspects, in particular, the intertemporal decomposition, the model 'robustness' (i.e. the semiparametric modelling, cf. the criticism of Heckman et al. 2008), the decomposition of the pay gap into a part due to endowments and another due to the returns of endowments which could be considered as 'pure' discrimination, etc. The next section introduces our semiparametric method (instead of a parametric) Mincer equation, for measuring sources and degrees of discrimination. "Data and Descriptive Statistics section" describes the data used and the trends in wage gaps between men and women through the estimation of density functions and descriptive statistics. "Regression Results section" presents the regression results obtained for wage gap in the mean and in the income percentiles. "Conclusions and Further Discussion" section concludes and discusses further issues like possible selection biases.

Methodology: A Semiparametric Decomposition of the Wage Gap

Typically, the analysis of wage discrimination employs either the wage decomposition attributed to Oaxaca (1973) and Blinder (1973) or the one introduced by Juhn et al. (1991, 1993), sometimes combined with the distributional approach of counterfactual wage quantiles (see e.g. Machado and Mata 2005). The difference between the two former mentioned methods lies basically in the way in which the discrimination is identified or, in other words, how wage differences are decomposed

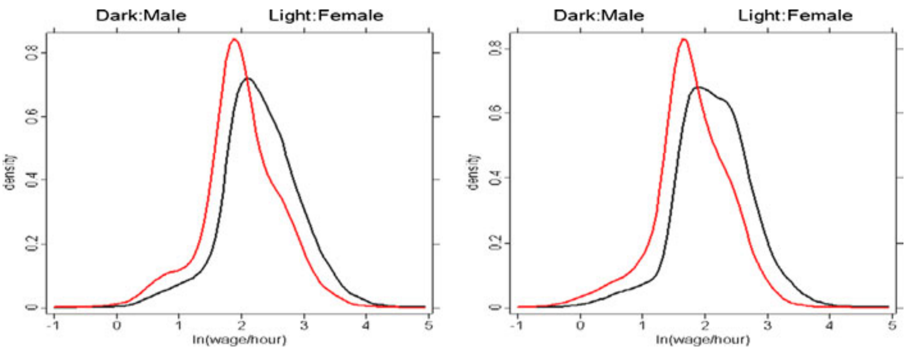

Fig. 1 Densities of ln(wage/hour) in 2002 (*left*) and 1995 (*right*)

and interpreted. In this study we have opted for the first, including extensions to quantile regression. We are interested in the development of the gender pay gap within a specific country, a question for which Oaxaca and Blinder's approach is easier to interpret.[1] Here, the gap between the wages of men and women is separated into two parts. The first is explained by the difference between the observed productive characteristics, the second lies in differences in the structure of the model, and is therefore not explained. This not explained part is usually considered as the wage discrimination by gender in the labor market.

A first limitation of the classical approach on which we focus is the specification of the wage equation. Although this problem is well known for the Mincer equation, see Heckman et al. (2008), Ichimura and Todd (2007), see also Lemieux (2006), when it comes to the analysis of wage discrimination, the existing approaches often use rather simple parametric specifications. Certainly, if the parametric model is chosen in an approximately correct form, the paradigm of the probability (Fisher 1922) provides estimators with good statistical properties. Unfortunately, the risk of incorrect specification is substantial, such that the statistical properties of the estimators obtained under the probability paradigm are often very poor, and the conclusions of the analysis might be incorrect. To get around this problem, some people use quadratic or higher order polynomials for age and experience, other add interaction or artificial cohort effect, etc. without justifying the (order of) polynomials. Furthermore, interactions of regressors are mostly ignored or quite specifically modeled. Instead of supposing that the regression function belongs to a function space characterized by a rather small number of parameters, we assume that it is an element in a bigger function space, like for example that one of sufficiently often differentiable functions. We use semi-parametric estimation techniques which combine flexibility and the possibility of modelling.

In this work we present an approach that offers an alternative to the classical specification of the regression equation. Instead of using Eq. 1, a most flexible form would be

$$\Lambda(W) = g(V) + \sigma(V)u$$

where W stands for wages, V are the explaining covariates, $\sigma(V)$ the standard deviation of the possibly heteroscedastic residual, and u is therefore the standardized residual. Further, g is an arbitrary nonparametric but smooth function, whereas Λ denotes a transformation such that we get additive errors, say U. It is generally accepted that for wage equations the logarithm is a reasonable choice for Λ. If part, say X, of the covariates $V=(X,Z)$ are dummy variables which (besides possible interaction) can have only a constant impact on wage, we should consider a semiparametric version of (3), namely

$$\ln W_i = X_i\beta + g(Z_i) + U_i, \tag{1}$$

[1] The alternative semiparametric extension of Juhn et al. (1991) is discussed in a note by Moral-Arce and Sperlich (2008). That method would for example be especially appropriate for country specific differences between the income distributions to later compare the gender pay gaps of these countries, see for example Blau and Kahn (1992, 2003).

where X_i is the set of observable qualitative characteristics (dummy variables) of that worker, Z_i further endowments, U_i an error term, β an unknown vector of parameters, and *g( )* a non-parametric function. With the superscript "w" referring to women and "m" to men, we consider the following system of equations:

$$\ln W_i^m = X_i^m \beta^m + g^m\left(Z_i^m\right) + U_i^m, \quad \ln W_i^w = X_i^w \beta^w + g^w\left(Z_i^w\right) + U_i^w, \tag{2}$$

but will supress in the following the individual's index *i* (unless it is indispensable) for the sake of notation. Then, like Oaxaca-Blinder, we express the difference between the two equations for year t as

$$\ln \overline{W_t^m} - \ln \overline{W_t^w} = \left(\overline{X_t^m} - \overline{X_t^w}\right)\beta_t^m + \left(\overline{g_t^m(Z_t^m)} - \overline{g_t^m(Z_t^w)}\right) + \left[\overline{X_t^w}\left(\beta_t^m - \beta_t^w\right) + \left(\overline{g_t^m(Z_t^w)} - \overline{g_t^w(Z_t^w)}\right)\right] \tag{3}$$

which allows us to obtain a first approximation of the wage discrimination that exists in certain countries. This total difference can be decomposed into two elements. The first two summands, also called "the explained part" are the wage gap due to personal characteristics (or endowments), which are measured for men and for women in the same way, e.g. age, experience, level of studies, etc. The third and fourth summands, the elements in brackets, represent the non explained part, and they reflect the wage difference which is caused by unobserved but different "wage structures" (or also called "differences in returns") between the two genders.

The adaptation of the model of Oaxaca-Blinder for the inter temporal comparison of wage disparities (compare Smith and Welch (1989) or Wellington (1993)) shows the advantage of using different wage structures for an arbitrary country in two different time periods. Starting with Eq. 3 for a country in the time period from 1995 to 2002, the difference in wage discrimination between these years can be written as:

$$\begin{aligned} D_{02} - D_{95} = {} & \underbrace{\left(\nabla\overline{X_{02}} - \nabla\overline{X_{95}}\right)\beta_{02}^m + \nabla\overline{g_{02}^m(Z_{02})} - \nabla\overline{g_{02}^m(Z_{95})}}_{\text{effect of observed endowments}} + \underbrace{\nabla\overline{X_{95}}\left(\beta_{02}^m - \beta_{95}^m\right) + \nabla\overline{g_{02}^m(Z_{95})} - \nabla\overline{g_{95}^m(Z_{95})}}_{\text{effect of observed returns}} \\ & + \underbrace{\left(\overline{X_{02}^w} - \overline{X_{95}^w}\right)\nabla\beta_{02} + \nabla\overline{g_{02}\left(Z_{02}^w\right)} - \nabla\overline{g_{02}\left(Z_{95}^w\right)}}_{\text{effect of women's endowments}} + \underbrace{\overline{X_{95}^w}\left(\nabla\beta_{02} - \nabla\beta_{95}\right) + \nabla\overline{g_{02}\left(Z_{95}^w\right)} - \nabla\overline{g_{95}\left(Z_{95}^w\right)}}_{\text{effect of differences in observed returns}} \end{aligned} \tag{4}$$

with the differential $\nabla\overline{X_t} = \overline{X_t^m} - \overline{X_t^w}$ for t=1995 and 2002 respectively, and similarly defined $\nabla\beta_t$ (which is typically assumed to be positive). Wage differences are expressed for each year by $D_t = \log\overline{W_t^m} - \log\overline{W_t^w}$. Furthermore, we use $\nabla\overline{g_t^m(Z_s)} = \overline{g_t^m\left(Z_s^m\right)} - \overline{g_t^m\left(Z_s^w\right)}$, and $\nabla\overline{g_t\left(Z_s^w\right)} = \overline{g_t^m\left(Z_s^w\right)} - \overline{g_t^w\left(Z_s^w\right)}$, etc. for t,$s$= 1995 and / or 2000 respectively. Recalling this decomposition, the differences between these two years are caused by four factors: The first reflects the differences which can be observed between the individual qualities (endowments) of men and women. The second reflects the differences in valuations (or returns) of those individual characteristics at the two different times. The third represents the differences in individual characteristics of women between the two years. The fourth characterizes the differences in valuations (or returns) that are observed for the same characteristics between men and women.

We are aware of criticism like that this would simplify the level of disaggregation of wage discrimination see Cotton (1988), Neumark (1988) or Oaxaca and Ransom

(1994). They propose to consider a discrimination-free wage structure, say for 2002, and look for example at the resulting differences for year t. The semiparametric analogue to their idea would be

$$\left(\overline{X_t^m}-\overline{X_t^w}\right)\beta_{02}^{*}+\left(\overline{g_{02}^{*}(Z_t^m)}-\overline{g_{02}^{*}(Z_t^w)}\right)+ \\ +\overline{X_t^m}\left(\beta_{02}^{m}-\beta_{02}^{*}\right)+\left(\overline{g_{02}^{m}(Z_t^m)}-\overline{g_{02}^{*}(Z_t^m)}\right)+\overline{X_t^w}\left(\beta_{02}^{*}-\beta_{02}^{w}\right)+\left(\overline{g_{02}^{*}(Z_t^w)}-\overline{g_{02}^{w}(Z_t^w)}\right) \tag{5}$$

where g_{02}^{*} and β_{02}^{*} represent a hypothetical wage structure for 2002 under the absence of discrimination. In practice one typically takes a weighted average of g_{02}^{w} and g_{02}^{m}, andβ_{02}^{w} andβ_{02}^{m} respectively. These, however, are certainly subjective choices; the maybe most common practice is to stick to the wage structure of men so that we are back in Eq. 3.

A more serious limitation is that our above introduced decomposition provides information only on the conditional mean, which implies that the size of the wage gap and the weights of the factors that make it up are constant throughout the wage scale. To avoid this problem, different method which analyse the wage gap by means of quantiles have been developed, see Buchinsky (1998), Martins and Pereira (2004), Sakellariou (2004), Garcia et al. (2001). Another way of avoiding the limitation indicated in this paragraph are the approached proposed by Di Nardo et al. (1996) or Butcher and Di Nardo (2002) looking directly at densities. We will propose below semiparametric extensions of Machado and Mata (2005). Other semiparametric approaches can be found in Cole and Green (1992), and Stengos and Sun (2008). For quantiles, the semiparametric specification of Eq. 1 for men at time point t is given by:

$$Q_\theta\left(\ln W_t^w\middle|X_t^m,Z_t^m\right)=X_t^m\beta_\theta^m+g_\theta^m\left(Z_t^m\right), \tag{6}$$

where $Q_\theta(\ )$ represents the quantile of order "θ" of the wage density function conditioned by X_t^m andZ_t^m. For the case of having a pure parametric specification, Koenker and Bassett (1978) and Buchinsky (1998) introduced a GMM estimator to obtain the regression parameters of interest. For the purely non-parametric setting readers are referred to the approaches taken by Yu and Jones (1998) and Hall et al. (1999), who proposed the estimation of the conditional distribution function. For the selection of the optimal bandwidth in the estimation procedure see Ruppert (1997). For semiparametric quantile regression Cole and Green (1992), and Stengos and Sun (2008) developed an estimation method for the parametric part as well as for the non-parametric one. Our quantile regression approach extends the concept of Machado and Mata (2005), who propose a decomposition process that combines quantile regression and bootstrap. First, quantile regression is used to obtain estimates of the conditional quantiles given in (6). The second idea involved is the theorem of probability integral transformation from elementary statistics: If the random variable U has a uniform distribution on [0,1], then $F^{-1}(U)$ has distribution F. Therefore, for any given X_i,Z_i and the random variable $\theta \approx U[0,1]$, $X_i^m\beta_\theta^m+g_\theta^m\left(Z_i^m\right)$ has the same distribution as $\ln W_\theta^w\middle|X_\theta^m,Z_\theta^m$. If X_θ^m,Z_θ^m is fixed and we take random variables from X, Z of the population, $X_i^m\beta_\theta^m+g_\theta^m\left(Z_i^m\right)$ has the same distribution as $\ln W_\theta^w$. The estimation process is formally represented by the following steps:

- Generate a random sample of size j with uniform distribution on [0,1]: $u_1, \ldots, u_j$
- Estimate for each gender j different coefficients and non-parametric functions of the quantile regression: $\widehat{\beta}^m_{ui}, \widehat{g}^m_{ui}(), \widehat{\beta}^w_{ui}, \widehat{g}^w_{ui}()$, i=1,....,j
- Generate for each gender a random sample of size j with replacement from the values of X, Z, denoted by $\left\{\widetilde{X}_i^m, \widetilde{Z}_i^m\right\}_{i=1}^j$ and $\left\{\widetilde{X}_i^w, \widetilde{Z}_i^w\right\}_{i=1}^j$
- Obtain $\left\{\ln \widetilde{W}_i^m = \widetilde{X}_i^m \widehat{\beta}^m_{u_i} + \widehat{g}^m_{u_i}\left(\widetilde{Z}_i^m\right)\right\}_{i=1}^j$ and $\left\{\ln \widetilde{W}_i^w = \widetilde{X}_i^w \widehat{\beta}^w_{u_i} + \widehat{g}^w_{u_i}\left(\widetilde{Z}_i^w\right)\right\}_{i=1}^j$ as a random sample of size j from the marginal distributions of ln W in accordance with Eq. 6
- Generate a random sample of the counter-factual distribution as follows: $\left\{\ln \widetilde{W}_i^{cf} = \widetilde{X}_i^m \widehat{\beta}^w_{u_i} + \widehat{g}^w_{u_i}\left(\widetilde{Z}_i^m\right)\right\}_{i=1}^j$ is a random sample of the wage distribution that will exist for women if all explanatory variables are distributed as they are for men.

Now the wage gap between the genders can be decomposed into the contribution of the coefficients and the contribution of the "covariates" using the technique of Machado and Mata (2005), who analysed changes in wage density. To simplify the comparisons of the decomposition of Oaxaca, we can decompose the quantiles in the wage distribution:

$$Q_\theta(\ln W^m) - Q_\theta(\ln W^w) = \left[Q_\theta\left(\ln \widetilde{W}^m\right) - Q_\theta\left(\ln \widetilde{W}^{cf}\right)\right] + \left[Q_\theta\left(\ln \widetilde{W}^{cf}\right) - Q_\theta\left(\ln \widetilde{W}^w\right)\right] + residual \tag{7}$$

The first term on the right hand side is the contribution of the parameters to the wage gap between the $\theta-$th quantile for men and the $\theta-$th quantile for women. The second is the contribution of the explanatory variables. The residual contains the simulation errors that appear when many simulations are carried out. Assuming that the quantile Eq. 6 is correctly specified, the error term disappears (asymptotically) in (7).

Data and Descriptive Statistics

The data used in this paper come from the *Structure of Earnings Survey* (SES) conducted by INE (the Spanish National Institute of Statistics), which employs a method similar to that used in surveys of wage structures in other European countries. The SES consists of a two-level sampling of Spanish companies (stratified sampling in the first stage for local units, and systematic sampling for the selection of workers at those units).[2] For 1995 there was a stratified sample of 161423 workers, and for 2002 the total sample was of 161370 including only those who had

[2] Note that for all estimation steps, including for the descriptive statistics, one has to account for the stratification by including the sampling weights.

full-time contracts and branches that existed in both years (losing only some still pretty small and recent branches). This survey collects information on non-self-employed workers who work at establishments with at least 10 workers and covers a wide range of private sectors (industry, construction, commerce, the hotel & catering business, transport, financial intermediation, etc.) excluding the primary sector. We once more emphasize that we concentrate on discrimination in the private sector for reasons discussed in the introduction, compare Aláez and Ullibarri (2001) also for the limitations arising.

Tables 1 (and 2, 3) summarize the variables we include in our model. As can be seen from there, the notion of "loyalty" or "tenure" (like the central endowment in the study of De la Rica et al. 2008) might be more appropriate than "experience" to describe what is actually measured. However, it must be noted that interruptions are

Table 1 Table of variables, in brackets the cluster used as reference group

Wage	Gross hourly earnings from employment
Ln(wage)	The natural logarithm of wage
Experience	Length of service in the actual enterprise, number of years
Age	Number of years of the employee
Intern. Market	Dummy: 1 if products are mostly sold outside Spain
(Loc-national Market)	Dummy: 1 if products are mostly sold in local or national market
Enterprise size 1	Dummy: 1 if between 16 and 25 employees in the enterprise
Enterprise size 2	Dummy: 1 if 25 or more employees in the enterprise
(Enterprise size 0)	Dummy: 1 if less than 16 employees
Ed. level 1	Dummy: 1 if high school or apprenticeship level studies are held
Ed. Level 2	Dummy: 1 if university studies are held
(Ed. Level 0)	Dummy: 1 if elementary studies are held
Long term	Dummy: 1 if contract is long-term
(Short term)	Dummy:1 if contract is short-term
C	Mining and quarrying industry
D	Manufacturing industry
E	Energy
F	Construction
G	Wholesale and retail trade, repair of motor vehicles
H	Accommodation and food service activities
I	Transportation and storage
J	Finance and insurance
(K)	Renting and other auxiliary activities
Professional	Professionals
Tecnic	Associate professionals and technicians
Clerk	Clerical support workers
Services	Service and sales workers
Operators	Plant and machine operators and assemblers
Nonskilled	Non skilled workers
(Managers)	Managers

Table 2 Descriptive statistics of ln(wage/hour) in 2002 and 1995

	2002		1995	
	Male	Female	Male	Female
Number of data	120317	41053	126743	34680
Min	−1.3325	−0.2794	−1.0076	−3.409
Max	5.3772	5.0052	4.864	4.2527
Mean	2.3676	2.1287	2.1884	1.8482
Median	2.2728	2.0204	2.1219	1.7665
Standard deviation	0.6344	0.5855	0.6138	0.6367

not counted, so this variable really reflects working time. Second, for Spain the difference between experience in a particular job and loyalty to the firm is often negligible since Spain has the lowest mobility of any industrialized country in the world. Third, in Spain loyalty is typically more important for salary purposes than what is commonly understood as "experience". Finally, note that by law (in force for the whole period of interest) women can return to their job (at least at the same level) after a maternity break of up to three years. To our knowledge, women in Spain often make use of this opportunity and typically do not use this break to change firms afterwards.

Note further that the list of covariates in Table 1 is the result of a previous model selection not documented here. A less typical variable in this kind of studies is *international market* but it turned out to be clearly significant. It reflects that the company is under the pressure of global competition but is certainly also correlated with size. As shown in Regression Results section, these covariates give always an R^2 larger than 50% for both genders and periods studied. Although we still obtain R^2s larger than 49% when excluding occupation dummies, we prefer to have them included on account of its importance for the Spanish gap decomposition; recall our discussion in the introduction and e.g. Aláez and Ullibarri (2001) or Fernandez et al. (2000).

Table 3 Descriptive statistics of endowments in 2002 and 1995

		2002		1995	
		Male	Female	Male	Female
Experience	Mean	9.886	7.748	12.237	9.4945
	Median	5.333	3.75	9.001	6.000
	St.dev.	10.399	9.000	10.2774	9.0723
Educ.level 1	Mean	0.5568	0.5448	0.5284	0.6584
	St.dev.	0.4968	0.4980	0.4991	0.4742
Educ.level 2	Mean	0.1718	0.3100	0.1209	0.1353
	St.dev.	0.3772	0.4625	0.3260	0.3421
Long term	Mean	0.7948	0.7722	0.7964	0.7296
	St.dev.	0.4037	0.4193	0.4026	0.4441

Some Basic Evidence

From Table 2 we see mainly 3 basic and expected facts. First, the proportion of female labor participation has increased substantially from 1995 to 2002 even though women are still more focusing on the public than on the private sector, partly because of less notable wage discrimination, see Ullibarri (2003) and Aláez et al. (2009), and partly because of the better compatibility with family; see also De la Rica et al. (2008) or Arulampalam et al. (2007). Second, no matter whether we look at the observed minimum or maximum wage, the mean or the median, thinking in percentages the gender gap in wages has seemingly decreased quite a bit. What we cannot see from this table is whether this is due to the (improved) endowments of women or whether this is due to a reduction of wage discrimination. Third, the standard deviation is not a good indicator; in general one would expect a smaller standard deviation for women than for men in 1995 and quite a similar one in 2002. But recalling that we have only people from the private sector in our sample, an easy explanation is that many of the female with middle-income preferred the public sector what may have risen the income dispersion for the remaining.

Table 3 exposes three quite interesting features of the changes in the Spanish labor force characteristics which are typical for Sothern European states, cf. Arulampalam et al. (2007), Aláez et al. (2009), Simón (2010) or the OCDE (2002). First, that "experience" has lowered just means that the originally very low mobility in Spain has increased from 1995 to 2002. Thereby, the here existing gender gap in experience decreased. Second, taking together the educational levels 1 and 2 we see that women have now outstripped men by far. It was about the same level for the highest degree in 1995, but already then about 20% higher for women in educational level 1. Third, women have caught up in respect to achieving long-term contracts which have an important positive impact on wage along basically all studies which are cited here and provided that information.

The Distributions of Log Wages

The study of wage data is based on two elements: first, descriptive statistics give information about position, dispersion and form of distribution, and second, a visual tool of the density functions applying nonparametric kernel estimation.

The logarithm of the hourly wage is used, calculated as the annual gross income divided by the number of hours worked in the year. The nonparametric density estimators consist of estimating f(x) without assuming that it belongs to a pre-established parametric family:

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) \tag{8}$$

with n being the total number of data available. Here, h is the width of the window and *K( )* is a kernel function. For more details, see Härdle et al. (2004).

Figures 1 and 2 show the densities of the wage per hour in logarithm for men and women separately for both years and their changes over time. Note that even though we face log wages, they can hardly be considered as being symmetric. Figure 1 for

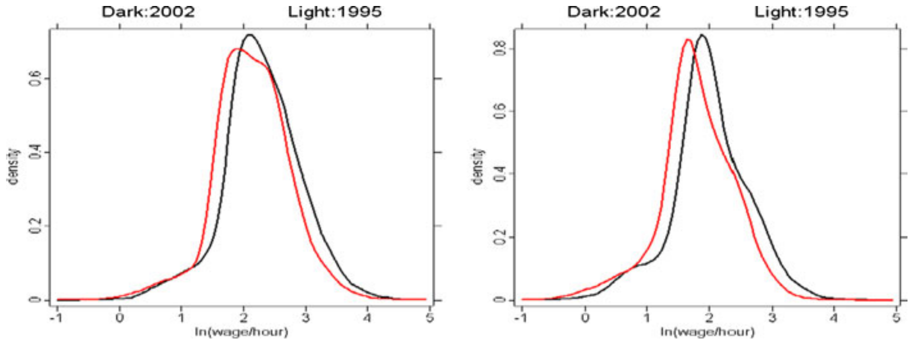

Fig. 2 Densities of ln(wage/hour), evolution:02 – 95 of male (*left*) and female (*right*)

1995 and 2002 seem to show that women tend to be located more on the left-hand side. Moreover, as the wage increases, their density always is below that of men. Concerning the trend in wages over time, Fig. 2 shows that there was a general increase in the wage for women, but no clear increase for men. The mode of the density for women is located more to the right in 2002 than it was in 1995, but for men it seems that no similar growth took place. Unquestionably, the mean increased between 1995 and 2002 with emphasis on the case of women. Note that this outcome and trend is also mentioned by most of the above cited articles but has not been that clearly demonstrated so far.

A Distribution Based Comparison of Wages

When no particular structure is considered, the most rudimentary, most flexible model that comes to mind is the joint distribution of variables in terms of their density. As the counterpart for the mixture of continuous and discrete data is more involved we restrict ourselves to illustrating the relationship between log wages and age and years of experience, respectively. To that end we use an extension of the nonparametric density estimator given in the previous section. If we have two variables (x,y) for n observations, density f(x, y) can be estimated by

$$\widehat{f}(x,y) = \frac{1}{nh_xh_y}\sum_{i=1}^{n} K_x\left(\frac{x-X_i}{h_x}\right)K_y\left(\frac{y-Y_i}{h_y}\right), K_x, K_y \text{ kernels with windows size } h_x, h_y \quad (9)$$

The joint density for log-wage and age (not shown) mainly revealed that the densities are much more peaked for women (at the age of about 30 and log-wage of about 2) than they are for men. Figures 3, 4, 5 and 6 show the results for the two-dimensional densities for log-wage (in the 3-D plots increasing from right to left for the sake of presentation) and experience in 1995 and 2002. The contours for men and women refer to the same density values. A cross line perpendicular to any point on the year axis gives a conditional distribution of the wage for a certain number of years. Keeping this in mind, contours are observed that run from north to south with a peak at about 0 to 4 years and about 2 for log-wages. For less than 5 years of experience all densities are much more spread (from negative to >3 log-wages), and skewed to the left, while for experience above 10 years the log-wage spread is much smaller ranging from 0.5 to 2 in 1995 and from 1 to 4 in 2002, respectively.

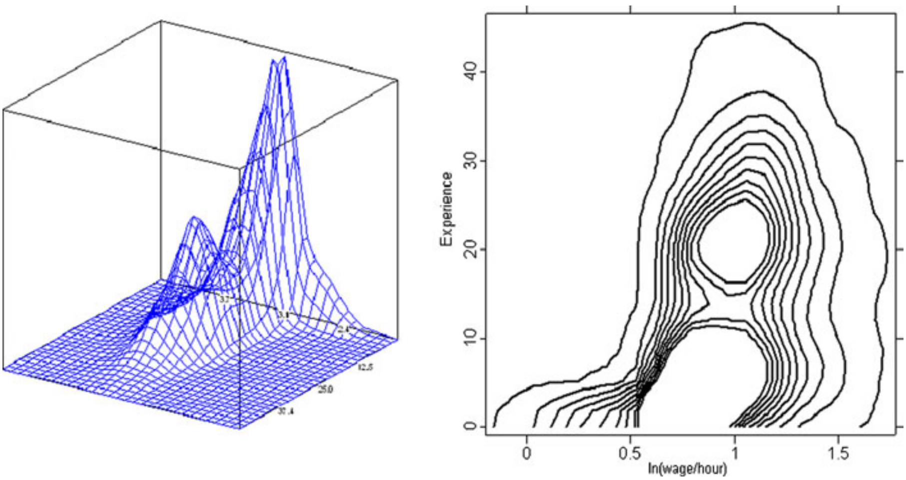

Fig. 3 Density of ln(wage/hour) and Experience in 1995 (male)

Different authors like e.g. De la Rica et al. (2008) argue that the shorter tenure of women caused by societal discrimination, leads to lower women's wages at the lower part of the distribution.[3] And indeed, especially when comparing the income distributions in 1995 for men vs. women with an experience of less than 10 years we see this hypothesis strongly confirmed (for experience 0–2 the density goes up and down much earlier and faster for women; for experience 2–10 we see a clear spread towards lower wages for women which is not there for men). The supposition that this gap would disappear for high tenure is less clear when checking the wage-scale (in accordance to Fig. 1). While these findings still seem to hold for 2002, they are definitely much less evident. Certainly, one might now argue that the here detected differences may be explained by other factors than discrimination. In order to test this, we now turn to the regression analysis of mean income and quantiles of the income distribution.

Regression Results

We will proceed in tow steps. We first carry out a mean regression analysis and the resulting wage gap decomposition, as introduced in Methodology: A Semiparametric Decomposition of the Wage Gap section. Afterward, we do the semiparametric quantile analysis which has also been introduced in that section.

Mean Regression

As already discussed, there are several possible causes of the aforesaid wage gap. On the one hand there is labor segregation, due to which women are paid poorly because they are employed in bad jobs. On the other hand it can be seen that employers offer

[3] More specifically, they argue that "*employers may use statistical discrimination in wage-setting in order to pay a lower proportion of the training cost for women than for men*".

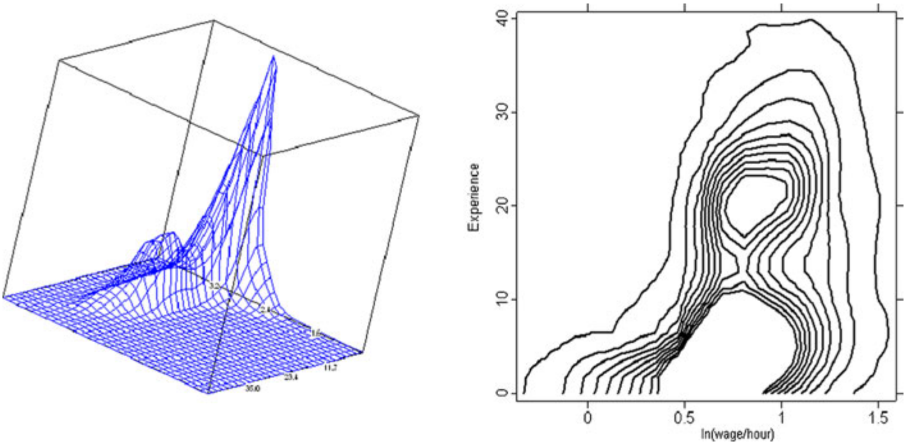

Fig. 4 Density of ln(wage/hour) and Experience in 1995 (female)

different wages for workers with quite similar abilities. This wage dispersion within a company is less accentuated when sectorial collective bargaining is stronger and when wage limits are fixed by the government or by agreements between unions and employers' associations. To account as much as possible for these different sources we included apart from the endowments of the individuals also occupational dummies and the sectors of professional activity.

Recall the mean regression based method introduced in Methodology: A Semiparametric Decomposition of the Wage Gap section. As all covariates are dummies except experience and age, only these latter two enter into the nonparametric part. Our first step consists of separately estimating Eq. 2 for both years. Table 4 shows the coefficient estimates. Most of them are significant. The variables that positively influence the wage of a worker are: belonging to a large company, having university education, and having a long-term contract. Most striking here is that right the returns to educational level went down significantly

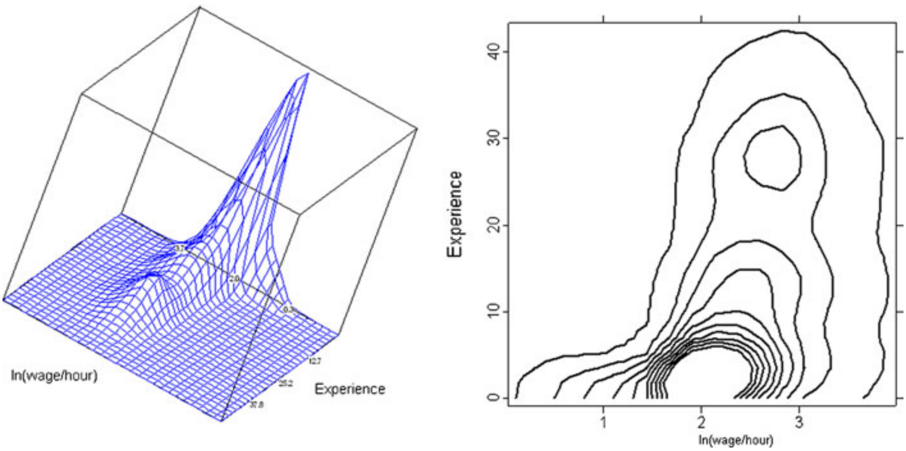

Fig. 5 Density of ln(wage/hour) and Experience in 2002 (male)

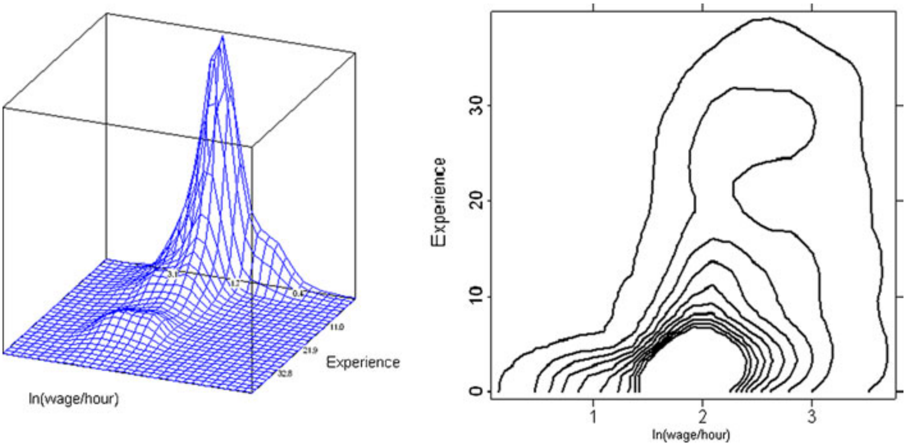

Fig. 6 Density of ln(wage/hour) and Experience in 2002 (female)

from 1995 to 2002, so right there where women have improved most.[4] In terms of sectors, energy supply (E) and the financial sector (J) influence wages strongly positively, whereas the influence of being in the hotel and catering business (H) is clearly negative. Note that all coefficients related to occupations show negative signs simply because 'manager' is our reference group. The ranking of these estimates are in accordance with what intuition would expect. As far as they are comparable with similar studies for Spain, they do not contradict to their findings, see for example Garcia et al. (2001), Simón (2010), Aláez et al. (2009), De la Rica et al. (2008) and others. Direct comparisons are hardly possible as they all use different years and samples, different covariates, and all work with (log-)linear models, typically not allowing for interactions.

The estimated joint impacts of experience and age, i.e. the nonparametric functions *g( )* are plotted in Figs. 7 and 8, separated by year and gender.[5] As *g( )* is a function from two to one dimension, it is presented as a 2-dimensional hyperplane on the joint support of the covariates *Experience* and *Age*. They show that both covariates have a strong though not linear impact on log-wages, independently of type and years. No major conclusions can be drawn for studies concerning the marginal impacts as interactions seem to dominate – something ignored in most of the parametric studies! Looking at men in 1995, we observe for people with low experience that wages increase log-linearly with age almost until 65. For people with higher experience, the impact of age is already flat for people older than 50, or even decreases. In contrast, experience has a clear positive impact for basically all ages below 65. In 2002, this has changed a bit. Now, log-wages seem to increase almost throughout (except for the oldest where experience has no significant impact once it is above 25) for both covariates with additional positive synergy effects, i.e. exhibiting a clear interaction. When we consider women in 1995, we observe a quite

[4] We are certainly aware of the quite abundant literature on over qualification which may justify this flattening of return to education without charging it to gender discrimination. However, most of the literature we know on this concerning Spain investigates several years before 1995. For a more recent study, see Budria and Moro-Egido (2006).

[5] The estimates of *g( )* were obtained with window sizes equal to 1.5 times Silverman's rule of thumb.

Table 4 Coefficient estimates and goodness of fit measure, Eq. 2

Year	2002				1995			
	Male		Female		Male		Female	
	Param	s.e.	Param	s.e.	Param.	s.e.	Param.	s.e.
Intern.market	0.0654	0.0033	0.0834	0.0053	0.0623	0.0032	0.0559	0.0061
Company size1	0.0979	0.0036	0.0670	0.0056	0.1055	0.0033	0.0596	0.0066
Company size2	0.2613	0.0037	0.2028	0.0054	0.2816	0.0034	0.1920	0.0066
Educ. Level 1	0.1138	0.0034	0.0515	0.0050	0.1047	0.0028	0.1016	0.0064
Educ. Level 2	0.2098	0.0051	0.1823	0.0063	0.3508	0.0052	0.2773	0.0099
Long term	0.3062	0.0037	0.3053	0.0050	0.3166	0.0041	0.2801	0.0070
C	0.2427	0.0087	0.1540	0.0282	0.1236	0.0096	0.0998	0.0334
D	0.0705	0.0042	0.0475	0.0056	−0.0054	0.0060	0.0212	0.0095
E	0.3155	0.0077	0.2743	0.0169	0.2970	0.0084	0.2297	0.0196
F	0.1498	0.0056	0.1089	0.0144	0.0439	0.0071	0.0788	0.0174
G	0.0437	0.0058	−0.0254	0.0067	−0.0255	0.0070	−0.0637	0.0108
H	−0.0811	0.0081	0.0027	0.0089	−0.1165	0.0086	−0.0403	0.0135
I	0.1258	0.0058	0.1143	0.0093	0.0786	0.0072	0.1530	0.0125
J	0.2768	0.0065	0.3208	0.0080	0.1611	0.0071	0.2520	0.0112
Professional	−0.2361	0.0082	−0.1991	0.0152	−0.2475	0.0071	−0.1423	0.0205
Tecnic	−0.4517	0.0075	−0.4742	0.0151	−0.3526	0.0062	−0.3342	0.0190
Clerk	−0.7097	0.0081	−0.6895	0.0153	−0.5882	0.0064	−0.5426	0.0184
Services	−0.7592	0.0091	−0.7637	0.0158	−0.6514	0.0080	−0.6283	0.0198
Operators	−0.7185	0.0079	−0.8528	0.0172	−0.6324	0.0063	−0.7982	0.0198
Nonskilled	−0.7407	0.0078	−0.8653	0.0162	−0.6365	0.0062	−0.7761	0.0192
R^2	0.5674		0.5853		0.5604		0.5596	

different impact picture than we did for men, especially that the impact of age is inversely U-shaped (except for very experienced and aged working women), and that for female workers experience caused already in 1995 increasing log-wages throughout. A maybe most striking finding is that for 2002 the general form of *g( )* is now similar to that of men (except maybe for very old women[6]) though on different scales; we detect the same general positive marginal effects of age and experience as well as the positive synergy (i.e. interaction) effect. Moreover, all slopes, bumps, and valleys we see in the log-wage surface for women in 2002 we can equally well detect for men in 2002, just on a different scale (what is partly expected according to the hypothesis of De la Rica et al. (2008) that the wage gap decreases for increasing tenure).

To study the past and actual wage discrimination, Table 5 shows the decomposition for 2002 and 1995 along Eq. 3. As can be seen there, the wage gap between men and women has reduced from 0.3404 in 1995 to still 0.2389 logarithmic points in 2002,

[6] There are very few old women with high experience (recall that this means tenure) which give a quite different shape to the curve at the upper right. Our general statements refer certainly to the mass of observations.

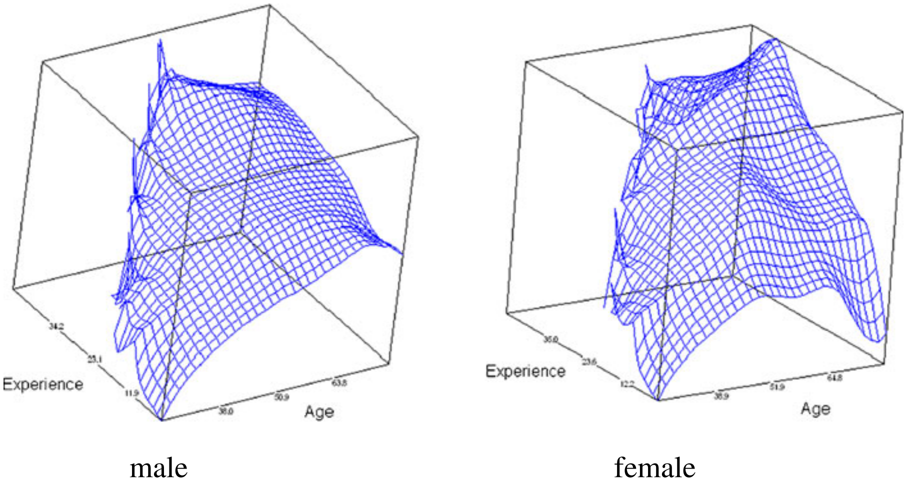

Fig. 7 Nonparametric function g of Eqs. 2 and 3. Year 1995

most of which is due to elements that cannot be explained (in neither of these two years). Moreover, in the explained part we see that as a result of their endowments women should actually earn about the same on average. In other words, the discrimination (unexplained part) is still evident (i.e. men earn about 23.8% more than identically qualified women). In view of all components, we must state that there was still a high degree of wage discrimination in Spain in 2002 although the observed total pay gap has reduced a lot compared to 1995.

Recall that Table 3 gave an ambiguous image of the development of endowments: while women had clearly outstripped men in 2002 concerning education, they did

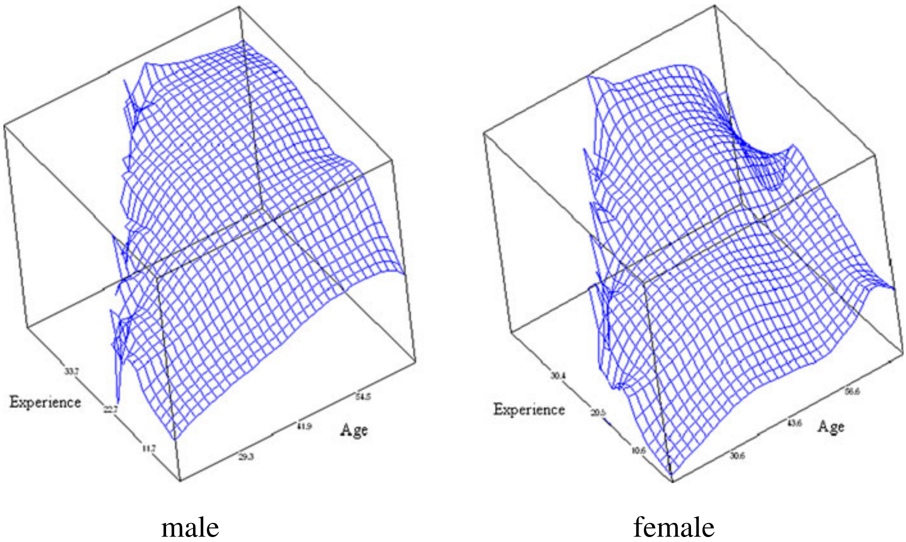

Fig. 8 Nonparametric function g of Eqs. 2 and 3. Year 2002

Table 5 Decomposition of Gender Pay Gap in Spain, 2002 and 1995, Eq. 3

Description	Estimates	
Year	2002	1995
Observed gap	0.2389	0.3403
Due to:		
Observed endowments	0.0009	0.0921
Unexplained	0.2380	0.2482

not succeed to improve their situation with respect to experience[7] and long term contracts, what is clearly in coherence with our discussion in the introduction and the statements of therein quoted articles (recall our discussion of child care or De la Rica et al. (2008) concerning social and statistical discrimination). To relate once again these descriptive statistics of endowments with wage, see Table 6. It shows the results of the decomposition of the trend in the wage gap from 1995 to 2002 along Eq. 4. One of the main conclusions that can be drawn is that the wage gap dropped by 0.101 logarithmic points between these years, from about 34% higher wages for men in 1995 to about 24% in 2002. At first glimpse this looks like good news, but if we look at the components, a remarkable fact is that most part of the reduction is due to the reduction of the differences in observed endowments between men and women and feminine endowments (with a value of −0.0972, and −0.0288 respectively). This implies that women considerably improved their characteristics (higher levels of study, more experience, acquisition of posts at better paying companies and with better contracts) compared to 1995 on the one hand, but also in comparison to men on the other hand. These results show that in the reference period women improved a great deal in relative terms compared to men and also in absolute terms according to an analysis of changes over time. But that means also that with respect to the valuation of these characteristics, say "observed returns", the situation of wage discrimination in Spain has unfortunately changed only slightly over these eight years.

Before coming to the analysis of income quantiles, we should briefly answer the question whether the use of the sophisticated semiparametric method was justified. We tested the validity of the alternative (log-)linear parametric analysis. We applied the nonparametric test of Härdle et al. (1998), see also Härdle et al. (2004). For both years and gender groups the parametric null hypothesis of a log wage equation had p-values clearly below 10% (based on bootstrap estimates of critical values). Notice that this may not just be due to nonlinearities but moreover due to the observed (compare Figs. 7 and 8) strong interactions in our regressions. In any case, this finding is important as it might (at least partly) be the reason for results deviating from other article's findings.

The Analysis Income Quantiles

A limitation of all above analysis is that it is based on average values, which prevents us from observing the trend in wage gaps along with the distribution of

[7] This is clear because in average women study much longer, see their educational level 2 in Table 3, time that rests from tenure.

Table 6 Decomposition of the gender pay gap in Spain, 1995–2002, Eq. 4

Description	Estimates
Observed change	−0.1014
Due to:	
Observed endowments	−0.0972
Observed returns	0.0015
Effect of women's endowments	−0.0288
Effect of differences in the observed return	0.0232

wages. Table 7 shows the estimated wage differentials for quantiles 0.1, 0.25, 0.5, 0.75 and 0.9 along the decomposition introduced in Eq. 7. All results are obtained using the estimation methods described in Methodology: A Semiparametric Decomposition of the Wage Gap section.

We see that for the highest wages the drop in the wage gap between 1995 and 2002 is higher than for the rest of the quantiles. This is easily seen in the 75th and 90th quantiles, which have a value of about 0.30 points in 1995 and about 0.20 in 2002. Note that, interestingly, in contrast to many other EU Member States (cf. Arulampalam et al. 2007), where the total observed gap increases monotonously with the quantiles, in Spain we observe a 'W-shape' for 1995, and a kind of asymmetric (mostly decreasing) U-shape in 2002, compare once again with Fig. 1a of De la Rica et al. (2008). The situation observed here, in which gender pay gaps are typically wider at the end and at the top of the wage distribution, is known as the "sticky floors" and "glass ceilings". The latter one has reduced a lot in 2002 compared to 1995 but the floors stay quite sticky for women, compare with del Rio et al. (2011) for 1995.

When analyzing inequality between men and women, this metaphor typically describes the barrier to further advancement once women have attained a certain level. From there on they see their male counterparts promoted while they are not. The "sticky floor" is simply the opposite scenario of the "glass ceiling". Here the gaps widen at the bottom of the wage distribution, an effect that has even aggravated in Spain from 1995 to 2002. Booth et al. (2003) define this as the situation where men and women with identical endowments might be appointed to the same pay scale, but the women are appointed at the bottom and men further up the scale. The explanation for these two effects can be rather complex: it involves the interplay of several factors like different reservation wages etc., cf. the arguments of De la Rica

Table 7 Evolution of the decomposition of the gaps by quantiles: Eq. 7

Quantiles	year	10th	25th	50th	75th	90th
Total observed gap	2002	0.3024	0.2282	0.2105	0.2012	0.2259
	1995	0.3497	0.2879	0.3496	0.3095	0.3326
Due to endowments	2002	−0.0476	0.0273	0.0554	0.0140	0.0675
	1995	−0.0044	0.0802	0.0187	0.0050	0.1180
Unexplained (due to returns)	2002	0.3500	0.2009	0.1551	0.1872	0.1584
	1995	0.3541	0.2077	0.3309	0.3045	0.2146

et al. (2008). Recall finally that due to the nature of our sample, the "sticky floor" effect observed here cannot be provoked by irregular immigrants or clandestine employment. However, it can certainly be influenced by regular immigration as suggested when thinking of the work of Butcher and Di Nardo (2002).

Several papers based on Spanish data from the mid or end nineties claim a U-shape, being widest for high wages. For 1995, our data and semiparametric model find rather a W-shape being widest for low income groups. A look at the decomposition of the total observed gap confirms our general findings above, but now separated for the different quantiles. The conclusion that most of the wage gap for the distribution is due to the valuation or "returns" to endowments in both years is even true for all quantiles. However, it seems that the gap has become significantly smaller for the higher wage groups. As for endowments, the trend in the gap has hardly changed, whereas the absolute values did a lot. For example, for the 10th quantile the value has fallen by 4 points, for the 25th and 90th quantile even by 5 to 6 points. Only for the middle-high wages (75th and 50th quantile) the gap increase from 1995 to 2002, and here even by more than 100%. Certainly, in average (pooled over the quantiles), these findings confirm our statements from above: women have improved their endowments but discrimination (gap due to returns) goes on. Interestingly, this phenomenon is quite unequally distributed over the income groups. Note that we have not found any comparable results in the other articles as they typically considered special partitions of their population or found at most U-shapes based on log-linear models without interactions.

Conclusions and Further Discussion

This paper examines the structure of the wage gap for men and women in the private sectors in Spain during the period from 1995 to 2002. Among other findings, a first simple descriptive analysis including nonparametric density estimation shows that, though wages seem to have increased for both genders in the period of reference, the growth in women's earnings is especially strong. Nevertheless, a notable pay gap is still existent. This, so far, coincides with the statements of similar but purely parametric studies.

As for the trend in the wage gap, this article introduces, justifies and applies different flexible semiparametric extensions of the regression based decomposition of Oaxaca-Blinder and its modifications. Our application of these new methods to Spain indicates a decrease in the gender pay gap over the period from 1995 to 2002, i.e. a period where many efforts have been made to combat discrimination including the pay gap, recall our discussion in the Introduction. However, this decrease is basically due to the fact that women improved their endowments in absolute and relative terms compared to men from 1995 to 2002. What in our opinion is especially striking here, is that women's educational level is even clearly above that of men, even if not so for long term contracts and experience.[8] While it was comparable at the beginning of the observed period, women outstripped men by far in higher education. Unfortunately, the returns changed accordingly such that

[8] See our discussion from above on the negative correlation between educational level and experience.

women's effort to catch up in this respect has not been rewarded accordingly. Moreover we can say that in general, with regard to the valuation (or pricing) of women's endowments, a major political challenge still remains. Recall that, is there a selection bias, we have only estimated the lower bound; the actual discrimination might even be more serious then, see also our discussion below.

Extending our new methods of decomposition to counterfactual quantile estimation (cf. Machado and Mata 2005), we have been able to study the trend in the wage gap along with the distribution of wages. Apart from some particularities of Spain (compared to other EU members) we find that discrimination is much more serious in rather low and rather high paid jobs than for jobs with median wages. While the U-shape in 2002 is consistent with most of the studies we found to this topic, we found wider gaps at the bottom than at the top (in 2002) and even a W-shape for 1995. However, this might be due to the selection of the particular population but is not necessarily just due to the semiparametric modelling. Having in mind that all the other articles use different populations, covariates and neglect interactions, they are hardly comparable with our results. Both, the gap due to endowments (going back to social discrimination) as well as the gap due to returns (wage discrimination) is quite inhomogeneously distributed over the quantiles and not monotonically rising or falling.

A possible criticism (see Garcia et al. 2001) concerns the problem of endogenous selection in labor force. If all relevant information can be summarized, say in a covariate vector T, and the error distribution including covariance structure were known, see Heckman (1980), then this correction could be reduced to the inclusion of a new element in our regression. The corresponding extension of our semiparametric model causes no problems for identification or estimation. That is, such an extension would generally be feasible for our semiparametric method though computationally demanding, especially if also the selection bias correction has to be done semiparametrically (else one needs assumptions similar to the Heckman (1980) or Tobit- approaches). However, the problem in practice is the 'if'. In our application we disregard this extension due to several problems. One reason is that we could not find a good instrument T. "Good" means that such an extension should not increase the variances of our estimates more significantly than changing the estimation results. Actually, in semiparametrics we often face the problem that minor corrections for possible biases increase the variance to an extent that in the end one loses out when looking at the mean squared error – the most relevant criterion for the quality of statistical inference. We also found that most of the papers on the gender pay gap did not correct for a possible selection bias: they either argue that it is negligible or they simply ignore it. A quite careful study in this direction was performed by De la Rica et al. (2008): they considered potential selection biases separately, partitioning their sample by gender, educational level, and age. For the considered year 1994 they found a significant selection bias only for young women with low education. Not surprisingly, this single significant selection bias was positive. A careful study of our above formulae reveals that in case of positive selection, all our numerical results on the gender pay gap can be considered as lower bounds of the actual gap. Consequently, a correction for this possible selection bias would not change our findings qualitatively but only quantitatively. This also holds true for our finding that the gender pay gap has decreased from 1995 to 2002, because the much higher female labor participation in 2002 indicates a smaller selection bias (compared to 1995) resulting in a sharper lower bound than that for 1995.

Acknowledgements We appreciated helpful discussion with R. Alaéz, M. Ullibarri, A. Alonso, and A. Madariaga (Instituto de la Mujer - Emakunde), C. Hundertmark and R. Ohinata, and the comments of an anonymous referee which helped us a lot to improve this article. The authors acknowledge financial support from FUNCAS (Fundacion de las Cajas de Ahorros) which published an earlier version as working paper (Nº 382/2008).

References

Aláez R, Ullibarri M (2001) Discriminación salarial por sexo: un análisis del sector privado y sus diferencias regionales en España. ICE, Revista de Economía 789:117–138

Aláez R, Longás JC, Ullibarri M (2003) Diferencias salariales en España: un análisis sectorial/regional. Investigaciones Regionales 3:5–24

Aláez R, Longás JC, Ullibarri M (2009) Visualising gender wage differences in the European Union. Gender, Work and Organization. Forthcoming

Arulampalam W, Booth AL, Bryan ML (2007) Is there a glass ceiling over Europe? Exploring the gender pay gap across the wages distribution. Ind Labor Relat Rev 60:163–186

Blau FD, Kahn LM (1992) The gender earnings gap: learning from international comparison. Am Econ Rev 82:533–538

Blau FD, Kahn LM (2003) Understanding international differences in the gender pay gap. J Labor Econ 21:106–144

Blinder AS (1973) Wage discrimination: reduced form and structural estimates. J Hum Resour 8:436–455

Booth A, Francesconi M, Frank J (2003) A sticky floors model of promotion, pay and gender. Eur Econ Rev 47:295–322

Buchinsky M (1998) Recent advances in quantile regression models: a practical guideline for empirical research. J Hum Resour 33:88–126

Budria S, Moro-Egido A (2006) Education, educational mismatch, and wage inequality: evidence for Spain. Econ Educ Rev 27:332–341

Butcher KF, Di Nardo J (2002) The immigrant and native-born wage distributions: evidence from United States censuses. Ind Labor Relat Rev 56:97–121

Cole TJ, Green PJ (1992) Smoothing reference centile curves: the LMS method and penalized likelihood. Stat Med 11:1305–1319

Cotton J (1988) On the decomposition of wage differentials. Rev Econ Stat 70:236–243

De la Rica S, Dolado JJ, Llorens V (2008) Ceilings or floors: gender wage gaps by education in Spain (with Erratum). J Popul Econ 21:751–778

del Rio C, Gradin C, Canto O (2011) The measurement of gender wage discrimination: the distributional approach revisited. J Econ Inequal 9:57–86

Di Nardo J, Fortin N, Lemieux T (1996) Labor market institutions and the distribution of wages, 1973–1992: a semiparametric approach. Econometrica 64:1001–1044

Fernandez M, Montuenga VM, Romeu AE (2000) Wage differentials and non-competitive behavior in the Spanish industry labor market, Documentos de traballo analise economica, nº 9, IDEGA

Fisher RA (1922) On the mathematical foundations of theoretical statistics. Phil Trans Roy Soc Lond, Series A 222:209–268

Garcia J, Hernandez PJ, Lopez-Nicolas A (2001) How wide is the gap? An investigation of gender wage differences using quantile regression. Empir Econ 26:149–167

Hall P, Wolff RCL, Yao Q (1999) Methods for estimating a conditional distribution function. J Am Stat Assoc 94:154–163

Härdle W, Mammen E, Müller M (1998) Testing parametric versus semiparametric modelling in generalized linear models. J Am Stat Assoc 93:1461–1474

Härdle W, Müller M, Sperlich S, Werwatz A (2004) Nonparametric and semiparametric models. Springer Verlag, Berlin

Heckman J (1980) Sample selection bias as a specification error. In: Smith JP (ed) Female labor supply: theory and estimation. Princeton University Press, pp 206–248

Heckman J, Lochner LJ, Todd PE (2008) Earnings functions and rates of return. J Hum Cap 2:1–31

Ichimura H, Todd PE (2007) Implementing nonparametric and semiparametric estimators. Handb Econ 6:5369–5468

Juhn C, Murphy K, Pierce B (1991) Accounting for the slowdown in black-white wage convergence. In: Kosters M (ed) Workers and their wages. American Enterprise Institute Press, pp 107–143

Juhn C, Murphy K, Pierce B (1993) Wage inequality and the rise in returns to skills. J Polit Econ 101:410–442

Koenker RW, Bassett G (1978) Regression quantiles. Econometrica 50:43–61

Lemieux T (2006) The Mincer equation thirty years after 'schooling, experience, and earnings'. In: Grossbard S. (ed) Jacob Mincer, A Pioneer of Modern Economics. Springer Verlag, pp 127–145

Machado J, Mata J (2005) Counterfactual decompositions of changes in wage distributions using quantile regression. J Appl Econ 20:445–465

Martins P, Pereira P (2004) Does education reduce wage inequality? Quantile regressions evidence from fifteen European countries. Labour Econ 11:355–371

Miller PW (2009) The gender pay gap in the US: does sector make a difference? J Lab Res 30(1):52–74

Moral-Arce I, Sperlich S (2008) A semiparametric Juhn, Murphy, and Pierce – Decomposition for the Spanish Gender Pay Gap: A methodological Note. Preprints of the Centre for Statistics ZfS-2008-04, Göttingen

Neumark D (1988) Employers discriminatory behaviour and the estimation of wage discrimination. J Hum Resour 23:279–295

Oaxaca RL (1973) Male-female wage differentials in urban labor markets. Int Econ Rev 14:693–709

Oaxaca RL, Ransom MR (1994) On discrimination and the decomposition of wage. J Econ 61:5–21

OCDE (2002) Perspectivas del Empleo, Informe de la OCDE. Ministerio de Trabajo y Asuntos Sociales, Madrid

Ruppert D (1997) Empirical-based bandwidths for local polynomial nonparametric regression and density estimation. J Am Stat Assoc 92:1049–1062

Sakellariou C (2004) Gender-earnings differentials using quantile regressions. J Lab Res 25:457–468

Simón H (2010) International differences in wage inequality: a new glance with European matched employer-employee data. Br J Ind Relat 48:310–346

Simón H, Sanromá E, Ramos E (2008) Labour segregation and immigrant and native-born wage distributions in spain: an analysis using matched employer-employee data. Span Econ Rev 10:135–168

Smith JP, Welch FR (1989) Black economic progress after myrdal. J Econ Lit 27:519–564

Stengos T, Sun Y (2008) The absolute health income hypothesis revisited: a semiparametric quantile regression approach. Empir Econ 35:395–412

Ullibarri M (2003) Diferencias salarias entre los sectores public y privado por género, escorialidad y edad. El caso de Espana. El trimestre economico 278:233–252

Wellington AJ (1993) Changes in the male/female wage gap, 1976–85. J Hum Resour 28:383–411

Yu K, Jones MC (1998) Local linear quantile regression. J Am Stat Assoc 93:228–237